(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 March 2002 (14.03.2002)

PCT

(10) International Publication Number
WO 02/20835 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB01/03970

(22) International Filing Date:
4 September 2001 (04.09.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0021667.1 4 September 2000 (04.09.2000) GB

(71) Applicant *(for all designated States except US)*: **GLAXO GROUP LIMITED** [GB/GB]; Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex UB6 0NN (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **XU, Chun-Fang** [GB/GB]; GlaxoSmithKline Research and Development, Discovery Genetics Europe, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY (GB). **PURVIS, Ian, James** [GB/GB]; GlaxoSmithKline Research and Development, Discovery Genetics Europe, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY (GB).

(74) Agents: **WOODS, Geoffrey, Corlett** et al.; J.A. Kemp & Co., 14 South Square, Gray's Inn, London WC1R 5JJ (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/20835 A2

(54) Title: GENETIC STUDY

(57) Abstract: Method of performing an association study comprising obtaining information about (i) genetic polymorphisms and (ii) phenotypes which are present in a sample population, performing a haplotype analysis on the genetic polymorphism information to deduce the haplotypes present in the population, and performing a statistical analysis to detect a correlation between the phenotypes and the deduced haplotypes, thereby determining whether there is an association between the polymorphisms and the phenotypes.

GENETIC STUDY

Field of the invention

The invention relates to a method of performing a genetic study.

Background of the invention

Association studies are performed to determine whether a particular region of the genome contributes to a phenotype. These studies are based on the detection of a correlation (association) between the presence of a polymorphism in the genomic region and a change in the phenotype. The phenotype may be predisposition or susceptibility to a particular disease or response to medication. Association studies can therefore be used to determine whether a particular gene is relevant in a disease or whether a particular polymorphism causes or contributes to the disease.

Summary of the invention

The invention provides a method of performing an association study comprising:

(a) obtaining information about (i) genetic polymorphisms and (ii) phenotypes which are present in a sample population,

(b) performing a haplotype analysis on the genetic polymorphism information to deduce the haplotypes present in the population, and

(c) performing a statistical analysis to detect a correlation between the phenotypes and the deduced haplotypes,

thereby determining whether there is an association between the polymorphisms and the phenotypes.

Brief Description of the Drawings

Figure 1 shows the SNPs of the *NAT2* gene.

Figure 2 shows the SNPs spanning 140 kb on chromosome X.

Detailed description of the invention

The invention provides a method for performing an association study in which a correlation between a computed haplotype and a phenotype is analysed. If such a

correlation is detected then this indicates that the region in which the haplotype occurs is able to affect (cause or contribute) to the phenotype.

The haplotype is defined by particular nucleotides at particular positions of the chromosome, and comprises at least 2 or more polymorphic regions (e.g. single nucleotide polymorphisms (SNP's) or microsatellites) typically in linkage disequilibrium with each other. Typically the haplotype comprises at least 2, 3, 5, 10 or more SNP's. The alleles of polymorphisms which are in linkage disequilibrium with each other in a population tend to be found together more often than expected on the same chromosome. Typically all of the remaining polymorphisms of the haplotype will be present on the chromosome at least 30% of the times, for example at least 40 %, 50%, 70% or 90, of the time any of the polymorphisms of the haplotype is present in the chromosome. The allele frequency of each of the polymorphisms in the haplotype generally varies from 1% to 50%. The frequency of the haplotypes defined by the polymorphisms will generally be 1% to 99%.

Thus any of the polymorphisms in the haplotype may also be present on chromosomes in the absence of the remaining polymorphisms of the haplotype, or in the form of a different haplotype. Generally at least 2, for example at least 5, 10, 100, 1000, 10^4, 10^5, 10^6, 10^8 or more polymorphisms are analysed in the study.

Polymorphisms which are in linkage disequilibrium are typically within 500kb, preferably within 400kb, 200kb, 100 kb, 50kb, 10kb, 5kb or 1 kb of each other, and thus typically the two polymorphisms in the haplotype which are most distant from each other will be within any of these distances from each other. Each of the polymorphisms is an insertion, deletion or substitution of a nucleotide. The polymorphism may be an A, T, C or G.

Typically the haplotype is typically in (or at least in the vicinity of) of a gene which expresses a product. The haplotype may or may not cause a different RNA (e.g. mRNA) or protein product to be expressed from the gene. The haplotype may be 5' to the coding region (e.g. within the promoter), in the coding region, in an intron or 3' to the coding region. The haplotype may stretch across more than one of these regions of the gene, or may stretch across a region which contains more than one gene.

The polymorphism information which is used to deduce the haplotypes present in the population is generally in the form of specifying the nucleotide (including a deletion and or insertion) and its position on the chromosome. The polymorphism

information typically details the polymorphisms present in at least 1 kb, 10 kb, 50 kb, 100 kb, 1 Mb, 10 Mb, 100 Mb or more of polynucleotide. These specified numbers of bases may or may not be fully contiguous, i.e. there may be sequence within the regions for which no polymorphism information is available. Typically the polymorphism information is from more than one chromosome (i.e. from more than one pair of homologous chromosomes), and in a preferred embodiment the association study is a genome wide study, so that the polymorphism information is from all the chromosomes. Thus the study does not need to be performed on a predetermined locus.

In a preferred embodiment the genetic information contains incomplete or no phase information for the polymorphisms. Thus the position of the polymorphisms is known but it is not known on which chromosome (of the two homologous chromosomes) the polymorphism is present. Therefore it is not possible to determine which alleles of the polymorphisms are present on the same chromosome. The haplotype analysis algorithms mentioned below may be used to deduce the phase information and thus determine the haplotypes present in the population.

The polymorphism information may be from known or unknown genes. Thus the polymorphisms which are analysed may be in genes which have not been fully defined (in terms of sequence or function).

The polymorphism information may be obtained from a database which contains the results of the genetic typing of the individuals being analysed in the study. The genetic typing comprises detecting the presence of a polymorphism in a region of the genome.

Typically the presence of the polymorphism is determined in a method that comprises contacting a polynucleotide of the individual with a specific binding agent for the polymorphism and determining whether the agent binds to the region of the polynucleotide which may contain the polymorphism, the binding of the agent to the polymorphism indicating that the individual carries the polymorphism. Generally the agent will also bind to flanking nucleotides on one or both sides of the polymorphism, for example at least 2, 5, 10, 15 or more flanking nucleotides in total or on each side.

In the case where the presence of the polymorphism is being determined in a polynucleotide it may be detected in the double stranded form, but is typically detected in the single stranded form.

The agent may be a polynucleotide (single or double stranded) typically with a length of at least 10 nucleotides, for example at least 15, 20, 30 or more polynucleotides. The agent may be molecule which is structurally related to polynucleotides that comprises units (such as purines or pyrimidines) able to participate in Watson-Crick base pairing.

A polynucleotide agent which is used in the method will generally bind to the polymorphism, and flanking sequence, of the polynucleotide of the individual in a sequence specific manner (e.g. hybridise in accordance with Watson-Crick base pairing) and thus typically has a sequence which is fully or partially complementary to the sequence of the polymorphism and flanking region.

Typically the agent is a probe. This may be labelled or may be capable of being labelled indirectly. The detection of the label may be used to detect the presence of the probe on (and hence bound to) the polynucleotide of the individual. The binding of the probe to the polynucleotide may be used to immobilise either the probe or the polynucleotide (and thus to separate it from one composition or solution).

In one embodiment the polynucleotide of the individual is immobilised on a solid support and then contacted with the probe. The presence of the probe immobilised to the solid support (via its binding to the polymorphism) is then detected, either directly by detecting a label on the probe or indirectly by contacting the probe with a moiety that binds the probe. In the case of detecting a polynucleotide polymorphism the solid support is generally made of nitrocellulose or nylon.

The method may be based on an oligonucleotide ligation assay in which two oligonucleotide probes are used. These probes bind to adjacent areas on the polynucleotide which contains the polymorphism, allowing (after binding) the two probes to be ligated together by an appropriate ligase enzyme. However the two probes will only bind (in a manner which allows ligation) to a polynucleotide that contains the polymorphism, and therefore the detection of the ligated product may be used to determine the presence of the polymorphism.

In one embodiment the probe is used in a heteroduplex analysis based system to detect polynucleotide polymorphisms. In such a system when the probe is bound to polynucleotide sequence containing the polymorphism it forms a heteroduplex at the site where the polymorphism occurs (i.e. it does not form a double strand structure). Such a heteroduplex structure can be detected by the use of an enzyme which single or double strand specific. Typically the probe is an RNA probe and the enzyme used

is RNAse H which cleaves the heteroduplex region, thus allowing the polymorphism to be detected by means of the detection of the cleavage products.

The method may be based on fluorescent chemical cleavage mismatch analysis which is described for example in PCR Methods and Applications 3, 268-71 (1994) and Proc. Natl. Acad. Sci. 85, 4397-4401 (1998).

In one embodiment the polynucleotide agent is able to act as a primer for a PCR reaction only if it binds a polynucleotide containing the polymorphism (i.e. a sequence- or allele-specific PCR system). Thus a PCR product will only be produced if the polymorphism is present in the polynucleotide of the individual. Thus the presence of the polymorphism may be determined by the detection of the PCR product. Preferably the region of the primer which is complementary to the polymorphism is at or near the 3' end of the primer. In one embodiment of this system the polynucleotide agent will bind to the wild-type sequence (in which the polymorphism is not present) but will not act as a primer for a PCR reaction.

The method may be an RFLP based system. This can be used if the presence of the polymorphism in the polynucleotide creates or destroys a restriction site which is recognised by a restriction enzyme. Thus treatment of a polynucleotide with such a polymorphism will lead to different products being produced compared to the corresponding wild-type sequence. Thus the detection of the presence of particular restriction digest products can be used to determine the presence of the polymorphism.

The presence of the polymorphism may be determined based on the change which the presence of the polymorphism makes to the mobility of the polynucleotide during gel electrophoresis, e.g. single-stranded conformation polymorphism (SSCP) analysis may be used. This measures the mobility of the single stranded polynucleotide on a denaturing gel compared to the corresponding wild-type polynucleotide, the detection of a difference in mobility indicating the presence of the polymorphism. Denaturing gradient gel electrophoresis (DDGE) is a similar system where the polynucleotide is electrophoresed through a gel with a denaturing gradient, a difference in mobility compared to the corresponding wild-type polynucleotide indicating the presence of the polymorphism.

The presence of the polymorphism may be determined using a fluorescent dye and quenching agent-based PCR assay such as the Taqman PCR detection system. This assay uses an allele specific primer comprising the sequence around, and

including, the polymorphism. The specific primer is labelled with a fluorescent dye at its 5' end , a quenching agent at its 3' end and a 3' phosphate group preventing the addition of nucleotides to it. Normally the fluorescence of the dye is quenched by the quenching agent present in the same primer. The allele specific primer is used in conjunction with a second primer capable of hybridising to either allele 5' of the polymorphism.

In the assay, when the allele comprising the polymorphism is present Taq DNA polymerase adds nucleotides to the non-specific primer until it reaches the specific primer. It then releases polynucleotides, the fluorescent dye and quenching agent from the specific primer through its endonuclease activity. The fluorescent dye is therefore no longer in proximity to the quenching agent and fluoresces. In the presence of the allele which does not comprise the polymorphism the mismatch between the specific primer and template inhibits the endonuclease activity of Taq and the fluorescent dye is not release from the quenching agent. Therefore by measuring the fluorescence emitted the presence or absence of the polymorphism can be determined.

The phenotype information may be obtained from a database that contains the results of measurement of phenotypes in the individuals in the study. The phenotypes may be discrete phenotypes (such as the presence or absence of a trait) or a continuous phenotype (represented by the magnitude of a trait). Typically the phenotype is related to a disease, such as the presence or absence of the disease. The phenotype may be presence, or magnitude of, a symptom. The phenotype may be susceptibility (predisposition) to the disease. The phenotype may be the response to medical or pharmaceutical treatment.

The disease may be one which substantially only has a genetic component, or one that also has an environmental component. Typically in the disease a particular gene product (such as a protein) is either not expressed, expressed at a reduced or elevated level, or expressed in a form which is functionally deficient. The disease may be one which is caused by a pathogen, such as a virus or bacterium.

The disease may be one which is caused by an immune response, such as an auto-immune disease. The disease may be a cancer. The disease may be caused by an abnormality in or damage to a particular organ, body sytem, tissue or cell type. The disease is typically caused by old age; stress, or a diet with high levels of lipid or

carbohydrate. The disease may be a neurodegenerative, neurological, cardiovascular, inflammatory, psychiatric respiratory or metabolic.

The phenotype may be a particular response to a therapeutic agent, such as a therapeutic effect or a deleterious side effect.

The individuals in the association study have different phenotypes, i.e. have differences in the trait which is being studied. Generally the study will comprise at least 50, 100, 500, 1000, 2000, 5000 or more individuals. In one embodiment the study contains less than 50%, such as less than 20% or 10% of first degree relatives. In a preferred embodiment the study is performed in the form of a case-control study. This study method generally comprises individuals who have a particular phenotype (cases) and individuals who do not have the phenotype (controls). In particular embodiments there may be more than one case group in the study. Other study formats include triads (parent and child), sib pairs, other small nuclear families (such as parent with two or more children) or families with extended relatives.

In a case-control study and other studies mentioned above which are investigating a continuous phenotype the cases may be defined as possessing the trait to the level of a threshold magnitude, i.e. more than or less than a particular level. Controls would then be defined as not possessing the trait to the extent of the threshold value. In the case of more than one case group more than one threshold value would be used.

The haplotypes may be computed by any suitable algorithm. The algorithm is generally able to deduce whether any of the types of haplotypes mentioned above are present. The algorithm is typically one which is able to predict the haplotype of unrelated individuals (and generally also the haplotype frequency) for which the polymorphism phase information is not available or is incomplete. Such an algorithm typically performs the following steps:

(i) assigning initial haplotype frequencies by sampling them at random from an appropriate distribution;

(ii) resolving unambiguous haplotypes by identifying individuals who are either homozygotes or single site-site heterozygotes across a defined region of genome (the "scan window"), and calculating their contribution to the corresponding haplotype classes;

(iii) for each multiple heterozygous (ambiguous) individual, calculating expected contributions to all haplotype types compatible with its genotype;

(iv) updating haplotype frequencies by counting each gamete type across individuals and dividing counts by twice the sample size; and

(v) iterating steps (ii) through (iv) until frequencies stabilise.

Suitable algorithms which infer haplotype frequency when only single-locus genotypes are scored are known in the art. In these situations, individuals that are heterozygous for more than one locus convey ambiguous information about the gametic phase and missing data techniques, such as the E-M algorithm, formalized by Dempster et al. (1977) Journal of the Royal Statistical Society B39, 1-38 are appropriate. Hill (1974) Heredity 33, 229-39 gave a cubic equation for the maximum likelihood estimate of a gametic frequency for the case of two loci and two alleles and proposed an iterative E-M solution.

Weir and Cockerham (1979) Heredity 42, 105-111 showed how the likelihood equations should be solved and all real roots examined for the case of two loci. Little and Rubin (1987) Statistical Analysis with Missing Data (Wiley and Sons, Inc.) provided a general description of the EM algorithm for the count (multinomial) data. Long et al. (1995) American Journal of Human Genetics 56, 799-810 discuss testing for linkage disequilibrium (LD) and higher order interactions, giving details for carrying out E-M computations in three-locus case. Chiano and Clayton (1998) Annals of Human Genetics 62, 55-65 outlined multiple locus E-M and testing strategies for a binary response (e.g. presence of disease). Slatkin and Excoffier (1996) Heredity 76, 377-83 studied E-M assumptions and behaviour of tests for disequilibrium on estimated frequencies.

In one embodiment the Bayesian statistical method described in Stephens et al (2001) Am. J. Hum. Gen, 68, 978-89 is used for haplotype reconstruction and haplotype frequency determination. According to this embodiment the haplotype analysis for a continuous trait is generally performed by:

selecting a subset of markers from the set of markers that may correlate with the continuous trait;

for each individual, obtaining a value of the continuous trait and a pair of alleles for each of the markers in the subset of markers;

for each individual, determining probabilities of haplotypes that are compatible with the alleles in the subset of markers; and

performing a regression on the probabilities of haplotypes that are compatible with the alleles in the subset of markers, for all the individuals, to determine correlations between the continuous trait and the haplotypes.

The method of performing regression may comprise

for each individual, sampling a first haplotype from the haplotypes that are compatible with the individual's set of alleles, to thereby define a second haplotype which is determined by the sampling of the first haplotype;

assigning the value of the continuous trait for the individual to both the first haplotype and the second haplotype, to thereby define a doubled sample size;

performing an analysis of variance by comparing average values of the trait among the sampled first and second haplotypes for all the individuals;

repeating the steps of sampling, assigning and performing to obtain a distribution of correlations of the continuous trait and the haplotypes; and

determining a value from the distribution that identifies a significance of the correlation.

The step of performing an analysis of variance may comprise:

defining a design matrix of first and second indicator values having two rows for each individual, where the second indicator value is associated with the first and second haplotypes and remaining positions in the design matrix are set to the first indicator value in the two rows; and

performing a regression on the design matrix, to thereby identify a correlation value between the value of the continuous trait and the first and second haplotypes.

The step of determining a value may comprise:

determining a median from the distribution that identifies a significance of the correlation.

The step of performing regression may comprise:

for each haplotype in the set, assigning a rank of significance;

for each individual, sampling a first haplotype from the haplotypes that are compatible with the individual's set of alleles, to thereby define a second haplotype which is determined by the sampling of the first haplotype;

assigning the value of the continuous trait for the individual to both the first haplotype and the second haplotype, to thereby define a doubled sample size;

performing a one degree of freedom regression on the ranks for the sampled first and second haplotypes for all the individuals;

repeating the steps of sampling, assigning the value of the continuous trait and performing a one degree of freedom regression to obtain a distribution of the correlation of the continuous trait and the haplotypes; and

determining a value from the distribution that identifies a significance of the correlation.

The step of performing a one degree of freedom regression may comprise:

defining a design matrix having two columns of the ranks of the first and second haplotypes and having two rows for each individual; and

performing a regression on the design matrix, to thereby identify a correlation value between the value of the continuous trait and the haplotypes.

The step of determining a value may again comprise:

determining a median from the distribution that identifies a significance of the correlation.

The step of performing regression may comprise:

relating the value of the continuous trait for each individual to a vector of estimated frequencies of all haplotypes; and

performing a multiple regression of the trait values on the vectors of estimated frequencies.

The step of determining may comprise performing an expectation-maximization.

In one embodiment the analysis comprises:

for each individual, determining probabilities of haplotypes that are compatible with the alleles in the subset of markers; and

performing a regression on the probabilities of haplotypes that are compatible with the alleles in the subset of markers, for all the individuals, to determine correlations between the continuous trait and the haplotypes.

The step of performing regression may comprise:

for each individual, sampling a first haplotype from the haplotypes that are compatible with the individual's set of alleles, to thereby define a second haplotype which is determined by the sampling of the first haplotype;

assigning the value of the continuous trait for the individual to both the first haplotype and the second haplotype, to thereby define a doubled sample size;

performing an analysis of variance by comparing average values of the trait among the sampled first and second haplotypes for all the individuals;

repeating the steps of sampling, assigning and performing to obtain a distribution of correlations of the continuous trait and the haplotypes; and

determining a value from the distribution that identifies a significance of the correlation.

The step of performing analysis of variance may comprise:

defining a design matrix of first and second indicator values having two rows for each individual, where the second indicator value is associated with the first and second haplotypes and remaining positions in the design matrix are set to the first indicator value in the two rows; and

performing a regression on the design matrix, to thereby identify a correlation value between the value of the continuous trait and the first and second haplotypes.

The step of determining a value may comprise:

determining a median from the distribution that identifies a significance of the correlation.

The step of performing regression may comprise:

for each haplotype in the set, assigning a rank of significance;

for each individual, sampling a first haplotype from the haplotypes that are compatible with the individual's set of alleles, to thereby define a second haplotype which is determined by the sampling of the first haplotype;

assigning the value of the continuous trait for the individual to both the first haplotype and the second haplotype, to thereby define a doubled sample size;

performing a one degree of freedom regression on the ranks for the sampled first and second haplotypes for all the individuals;

repeating the steps of sampling, assigning the value of the continuous trait and performing a one degree of freedom regression to obtain a distribution of the correlation of the continuous trait and the haplotypes; and

determining a value from the distribution that identifies a significance of the correlation.

The step of performing a one degree of freedom regression may comprise:

defining a design matrix having two columns of the ranks of the first and second haplotypes and having two rows for each individual; and

performing a regression on the design matrix, to thereby identify a correlation value between the value of the continuous trait and the haplotypes.

The step of determining a value may comprise:

determining a median from the distribution that identifies a significance of the correlation.

The step of performing regression may comprise:

relating the value of the continuous trait for each individual to a vector of estimated frequencies of all haplotypes; and

performing a multiple regression of the trait values on the vectors of estimated frequencies.

The step of determining may comprise performing an expectation-maximization.

Further statistical tests for association of haplotype frequencies with continuous traits are discussed below. Such tests can have improved sensitivity because haplotypes can be stronger predictors of traits when there is lack of recombination. Conditions for asymptotic equivalence of standard regression-based methods with methods that "double the same size" will be described in the case of known haplotypes. These models then will be extended to the case of inferred haplotypes. Haplotype frequencies can be estimated through expectation-maximization (E-M), and each individual in a sample is expanded into all possible haplotype configurations with corresponding probabilities.

In the analysis a subset of markers may be selected from the set of markers that may correlate with the continuous trait. The selection of a subset of markers may be determined empirically and/or theoretically based on available literature, studies and/or other techniques. The selection of a subset of markers that may correlate with the continuous trait is well known to those having skill in the art and need not be described further herein.

For each individual, a value of the continuous trait and the pair of alleles for each of the markers in the subset of markers is obtained. The obtaining of a value of the continuous trait and the pair of alleles for each of the markers may be obtained through clinical trials or other studies that may involve a control group and a sample group. The obtaining a value of a continuous trait and a pair of alleles for each of the markers in the subset of markers is well known to those having skill in the art and need not be described further herein. The table below illustrates an example of data that may be thus obtained.

Individual	Continuous Trait	Marker 1	Marker 2	...	Marker L
1	1.01	11	12	...	13
2	-0.5	12	22	...	11
3	2.6	11	11	...	22
4	0.2	22	11	...	12
5	2.0	13	12	...	12
⋮	⋮	⋮	⋮	...	⋮
N	-1.3	22	12	...	21

For each individual, the value of the continuous trait and the allele numbers for markers 1-L are obtained.

For each individual the probabilities of haplotypes that are compatible with the alleles in the subset of markers is determined. Then, a regression is performed on the probabilities of haplotypes that are compatible with the alleles in the subset of markers, for all of the individuals, to determine correlations between the continuous trait and the haplotypes.

For example a first haplotype from the haplotypes that are compatible with the individual set of alleles is sampled from the probability distribution determined, to thereby define a second haplotype which is determined by the sampling of the first haplotype. The value of the continuous trait for the individual is assigned to both the first haplotype and the second haplotype, to thereby define a doubled sample size. An analysis of variance may be performed by comparing average values of the trait among the sampled first and second haplotypes for all the individuals. These operations are repeated a sufficient number of times, to obtain a distribution of correlations of the continuous trait and the haplotypes. When all the haplotypes have been processed, a value is determined from the distribution that identifies a significance of the correlation.

An analysis of variance may be performed by defining a design matrix of first and second indicator values (such as 0 and 1) having two rows for each individual, where the second indicator value is associated with the first and second haplotypes and remaining positions in the design matrix are set to the first indicator value in the two rows. A regression in then performed on the design matrix, to thereby identify a correlation value between the value of the continuous trait and the first and second haplotypes.

In Sasieni ((1997) From Genotype to Genes: Doubling the Sample Size, Biometrics 53, 1253-61) allelic versus genotypic tests for the case-control design and

bi-allelic markers were studied. As described therein, a genotypic test for association can operate on a 2 x 3 contingency table of individuals, classified according to their genotypes and the affection status. The total count of such a table is n. An allelic test would operate on a 2 x 2 table of allele counts versus affection status. Thus, each individual would contribute two alleles to the table, and the total count becomes 2n. The test implicitly assumes that the allele counts are binomially distributed, and thus may require that the population is in Hardy-Weinberg Equilibrium (HWE). Sasieni described that the Armitage's trend test addresses essentially the same question, however it does not "double" the data, and therefore can be applied to samples from non-randomly mating populations. Sasieni also provided explicit expressions for odds ratios comparing heterozygous and homozygous cases and argued that the genotypic test is sometimes a better choice, since it allows to test genotypic effects not explained by alleles, or "dominance deviations".

It may not be clear which test should be preferred for the multi-allelic markers, and especially haplotypes, because, for a marker with L alleles, the number of possible genotypes is L(L+1)/2, which may lead to large degrees of freedom tests and sparse tables. The allele-based test, on the other hand, will have L categories. Thus, sparseness may be less of a problem. An intrinsic assumption of allelic tests is that the response can be explained by the allelic "main effects". Then, certain situations, e.g. the two allele case when both homozygotes have the same effects, different from the effect of the heterozygote, may not be detected. Nevertheless, allele-based tests still are sensitive in many cases, if not uniformly most powerful.

A justification for the data-doubling in the case of continuous traits, now will be provided. Let X denote a gamete that may take any one of L values; X=j denotes that the gamete takes the jth allelic value ("j" is a label for the particular allele). The gametes may be single locus, in which case the values of X are called alleles. When X is multi-locus they are called haplotypes. Individual i has two gametes, X_{i1} and X_{i2}, and the genotype of individual i is denoted (X_{i1}, X_{i2}). Individual i also has an associated phenotype, Y_i.

Consider the 2n-dimensional linear model relating responses to gametic phase:

$$Y_2 = A\alpha + \varepsilon \qquad (1)$$

Here $A'=(A_{11}, A_{12}, A_{21}, A_{22},\ldots, A_{n1}, A_{n2})$, where A'_{ij} is the is the 1 x L allele indicator vector for gamete j of subject i. For example, if $X_{ij} = 2$, then $A'_{ij} = (0\ 1\ 0\ \ldots\ 0)$,

indicating that gamete j has allelic class 2. It will be understood that labeling of gametes as either 1 or 2 within an individual may be arbitrary. Let the elements of Y_2 denote corresponding phenotypic observations: $Y_2' = (Y_1, Y_1, Y_2, Y_2, \ldots, Y_n, Y_n)$, so that the data are doubled.

Equation (1) is an Analysis of Variance (ANOVA) model relating response to allele class. One may test for no affect of allele class on phenotype using the F test distribution with degrees of freedom L-1 and 2n-L: here

$$F = \{SSA/(L-1)\}/\{SSE/(2n-L)\},$$

with

$$SSA = Y_2'(A(A'A)^{-1}A' - J_{2nx2n}/(2n))Y_2$$

and

$$SSE = Y_2'(I_{2n}-A(A'A)^{-1}A')Y_2$$

and where J_{axb} denotes the (a x b) matrix of 1's and I_a denotes the (a x a) identity matrix.

The F test may be suspect because the response variable has been doubled. In other words, it may appear like "cheating" to artificially double the sample size. However, the F statistic is equivalent to that of the following n-dimensional regression model, and the data doubling is therefore valid.

An alternative model to Equation (1) with similar asymptotic properties and well-known finite-sample properties now will be described. The model is an n-dimensional regression model

$$Y = D\beta + \varepsilon \qquad (2)$$

where Y_i = trait value for individual i, $D' = (D_1, D_2, \ldots, D_n)$, $D_i' = (D_{i1}, D_{i2}, \ldots, D_{iL})$, and where

$$D_{ij}\begin{cases}2 & \text{if ith individual is homozygous for allele j}\\ 1 & \text{when ith individual is heterozygous including allele j}\\ 0 & \text{otherwise}\end{cases}$$

There is some correspondence between Equations (1) and (2) in that $A_{1i} + A_{2i} = D_i$. However, it is also clear that Equation (2) may have the usual validity (or lack thereof, in cases of lack of fit) of standard regression models, whereas Equation (1) may seem outrageous since the observations are simply doubled. Nevertheless, it will be shown that these models can produce equivalent F statistics when HWE holds.

To understand potential lack of fit, note that Equation (2) may correspond to that of Weir et al. (1977) Two-locus theory in Quantitative Genetics, Proceedings of the International Conference on Qunatitative Genetics pp247-69 and Nielsen et al. (1998) Am. J. Hum. Genetics 63, 1531-40, in the case of no "dominance" effects. Specifically, these publications assume the mean phenotypic response for genotype (j,k) is

$$\begin{aligned} \mu_{jk} &= \mu + \alpha_j + \alpha_k + d_{jk}, \quad \text{where} \\ \Sigma_j p_j \alpha_j &= 0, \quad \text{and} \\ \Sigma_j p_j d_{jk} &= \Sigma_k p_k d_{jk} = 0, \end{aligned} \tag{3}$$

the p_j denoting population allele frequencies. Equation (2) is exactly equation (3) with $d_{jk} \equiv 0$.

Equation (3) may lack sensitivity in cases of dominance effects $d_{jk} \not\equiv 0$). However, in the case of larger L (anticipating the case where the alleles are multi-locus haplotypes), the test for $\{H_0 : \alpha_i \equiv 0 \text{ and } d_{jk} \equiv 0\}$ may lose power because of the large numerator degrees of freedom (L(L+1)/2 - 1). In this case, the additive Equation (2) may be preferable despite possible lack of fit. In the additive model, the relation $2\beta_j = \mu + 2\alpha_j$ applies and in this case the test of $H_0 : \beta_1=\beta_2=\ldots=\beta_L$ is an L-1 numerator degrees of freedom test of no effect of allele on response. The F test uses:

$$F_1=\{SSA_1/(L-1)\}/\{SSE_1/(n-L)\},$$

with

$$SSA_1=Y'(D(D'D)^{-1}D'-J_{nxn}/n)Y$$

and

$$SSE_1=Y'(I_n-D(D'D)^{-1}D')Y$$

The statistics F and F_1 may be equivalent under the null hypothesis and F_1 dominates F under the alternative, under HWE. In particular, it will be shown below that:

$$F_1 - F \rightarrow_p 0 \tag{4}$$

when genotype has no effect on trait, and that:

$$F_1 / F \rightarrow_p c > 1 \tag{5}$$

when $V_\alpha = \Sigma p_j \alpha_j^2 - (\Sigma p_j \alpha_j)^2 > 0$. These results establish validity of the 2n-model Equation (1) under the null hypothesis. However, because the n-model Equation (2)

does not require HWE assumption and has an asymptotically larger F value under the alternative, it may be preferred. Nevertheless, the equivalence of the two approaches is useful for developing methodologies in the more complicated situations where alleles (specifically, multi-locus alleles, or haplotypes) are unobservable, as will now be shown.

Determining Probabilities

Operations for determining probabilities of haplotypes that are compatible with the alleles in the subset of the markers now will be described in detail.

In Equations (1) and (2), the "alleles" can denote multi-locus haplotypes rather than single-locus alleles. In this case the parameter α_j refers to the main effect of haplotype j. The haplotypes are generally unobservable, and therefore missing data methods may be used for their estimation. Consider two basic types of models. One is like the ANOVA model Equation (1), but where the A_{ij} are generated at random from a distribution inferred through the observed single-locus genotypes, then results are averaged over random haplotype generations. The second basic type is like the regression model Equation (2), where instead of using actual haplotype frequencies (0,1,2) for person i, the expected haplotype frequencies (given the observed single locus genotypes) are used.

Expectation-Maximization (E-M) techniques for inferring haplotype frequencies have been described previously. See the above-cited publications by Hill, Weir (1996), Long et al., Slatkin et al. and Chiano et al.. However, a review of E-M will be provided because it can be used in embodiments of the invention.

First, initial haplotype frequencies f are sampled from a symmetric Dirichlet distribution, Dir(γ, γ, ..., γ), where the dimension of the distribution is determined by the number of haplotypes compatible with the genotypes in the sample. Values of $\gamma>1$ result in more similar initial frequencies. The value used is $\gamma=1$ (multivariate uniform distribution). Each person's multilocus genotype (of L loci) is expanded into a set of possible haplotypes. This may be done by generating ($2^L/2 = 2^{L-1}$) vectors of zeros and ones, indicating whether the first or the second allele of an individual at the current locus should be taken. (This does not assume that loci are bi-allelic.) If L=3, the set is 000, 010, 100, and 110. This "primary" set of haplotypes assumes its complement, 111, 101, 011, and 001, and the maximum overall number of haplotypes

compatible with a given genotype is 2^L. The above expansion only needs to be performed once, at the E-M initialization stage. Next, haplotype frequencies (real values) are mapped to corresponding vectors of possible haplotypes (vectors of integers) for each possible haplotype in a sample. Long at al. used L-dimensional array indices for the mapping. However, the mapping may be more conveniently implemented through associative arrays, such as generic "map" from the C++ Standard Template Library. This can make the algorithm completely general with respect to the value of L.

Denote frequencies of haplotypes from the primary and complementary sets by f, and f^c. Per-subject probabilities are then updated, under assumption of HWE, as follows:

$$P_i = \sum_{j \in G_i} f_j f_j^c$$

where $|G_i| = 2^{L-1}$. The sample log-likelihood is:

$$\log l = \sum_{i=1}^{n} \ln P_i$$

summing over all n individuals. Then the f, and f^c frequencies are updated. For each f_i, and f_i^c the updates f_i' and $f_i^{c\prime}$ are

$$f_i' = \frac{1}{2n} \sum_{j=1}^{n} \frac{m_{ij} f_i f_i^c}{P_j}$$

$$f_i^{c\prime} = \frac{1}{2n} \sum_{j=1}^{n} \frac{m_{ij}^c f_i f_i^c}{P_j}$$

where m_{ij}, m_{ij}^c are the numbers of times that the haplotype i (or its complement) was counted as compatible with j-th individual's genotype. For example, if L=2, the 0-1 expansion is 00, 10 for a primary set, and 11, 01 for the complement. If the j-th individual had A_1/A_1 B_1/B_3 genotype, that would translate into A_1B_1, A_1B_1 haplotypes for the primary set, and A_1B_3, A_1B_3 for the complementary set. For updating the haplotype A_2B_3, then m_{ij} is zero, but if the haplotype is A_1B_3, then m_{ij} is equal to two.

The updating process (iteration) continues until the difference between subsequent sample log-likelihoods is sufficiently small. Several re-runs, starting from the random initialization of the haplotype frequencies may be needed to avoid local convergence, and the run with the highest log-likelihood should be taken. Estimated haplotype counts are given by the final values of 2nf, $2nf^c$.

Regression

The performance of regression on the probabilities of haplotypes that are compatible with the alleles in the subset of markers now will be described in detail.

As noted previously, an F test and its p-value are available in the case of known haplotypes. However, the actual haplotypes generally are not available, only an estimate of the probability distribution of their values. Thus, the distribution of likely values of the "true p-value" is estimated and a "likely" p-value may be picked as follows:

1. For each individual i=1,...,n, a pair of compatible haplotypes is randomly drawn from the distribution specified by the compatible haplotype frequencies;
2. A model specified by (1) is formed, and a test statistic (F) for the importance of including the genotype is calculated;
3. Operations 1 and 2 are repeated many times, and the p-value from each run is saved; then
4. The final p-value is given by the median of the distribution of p-values.

It is possible to greatly increase power of the test if some of the vector haplotype indicators A_{i1}, A_{i2} can be replaced by corresponding scalar rank scores R_{i1}, R_{i2} (in order of "importance"), based on prior tests or biological knowledge. In that case the model is formed as

$$Y_i=\mu+R_{i1}\beta+\epsilon_{i1}$$

$$Y_i=\mu+R_{i2}\beta+\epsilon_{i2}$$

This can concentrate the effect into a test with a single degree of freedom, and can have much greater power when the rank scores are chosen well.

An alternative is to perform a multiple regression, based on n observations instead of 2n, directly on the set of per-person expected haplotype frequencies. This is motivated by Equation (2), where the traits are regressed on the observed frequencies. If all elements in the matrix D in Equation (2) are divided by two, then they can be considered as probabilities for the individuals to have a particular allele. In the single-locus case, the identification of alleles may be certain, and so 0, 0.5, and 1 generally are the only values possible. In the case of E-M inferred haplotypes the corresponding model is:

$$Y_i = \sum_{j=1}^{L} f_{ij}\beta_j + \varepsilon_i$$

Frequencies for haplotypes incompatible with the ith individual's single-locus genotypes are set to zero. Also, haplotypes with expected counts that are less than one are removed from consideration. The usual F test of $H_0 : \beta_1=\ldots=\beta_L$ provides a test for effect of haplotypes on trait, and individual haplotype effects are tested as $H_0 : \beta_1=0$. The test can be made more robust by permuting the vector $(Y_1,\ldots,Y_n)$ independently of the haplotype frequency data. The final p-value is the proportion of permutations that yield an F-statistic p-value that is no larger than the original F-statistic p-value. However, the asymptotic p-values are valid in most situations.

Additional theoretical details now will be provided. In particular, 2n may be corresponded with n by labeling each of the n individuals by i, wherein $i \in S = \{1, 2, \ldots, n\}$. Partition S disjointly and exhaustively into sets N_{jk} so that $S = U_{j=1}^{L} U_{k=j}^{L} N_{jk}$, where N_{jk} is the set of individuals with genotype (j, k). Define $N_{kj}=N_{jk}$ and let $n_{jk}=|N_{jk}|$ denote the number of individuals out of n having genotype (j, k).

Consider a sequence of samples for which $n \rightarrow \infty$, under random sampling from an infinitely large population of individuals. In this case:

$$n_{jk}/n=p_{jk}+o_p(1) \qquad (6)$$

where p_{jk} is the population proportion of individuals with genotype (j, k), and where "$o_p(1)$" denotes a term that converges to 0 in probability. If the population is in HWE,

$$p_{jj} = p_j^2 \qquad (7)$$

and

$$p_{jk}=2p_jp_k \qquad (8)$$

where p_j denotes the population proportion of alleles with type j.

In the sample of n individuals there are 2n gametes. Of these 2n, there are

$$n_{Aj} = n_{jj} + (n_{j1} + \ldots + n_{jj} + \ldots + n_{jL})$$

gametes having allele j. Under random sampling and HWE,

$$n_{Aj}/(2n)=p_j+o_p(1) \qquad (9)$$

Equations (6)-(9) concern the behavior of the n_{jk} and the p_{jk}. Now consider the Y_i under embodiment 3. Assume that $Y_i=\mu_{jk}+\epsilon_i$ in set N_{jk}, where the ϵ_i are independent with $Var(\epsilon_i)=\sigma^2>0$.

Note that the sum of the Y_i corresponding to occurrences of allele j is the "allelic sum"

$$Y_{Aj}^{+} = 2\Sigma_{i\epsilon Njj} Y_i + \Sigma_{k\neq j}\Sigma_{i\epsilon Njk} Y_i,$$

and the corresponding "allelic average" is

$$\overline{Y}_{Aj} = Y_{Aj}^{+} / n_{Aj}$$

Using this model, it can be seen that:

$$\overline{Y}_{Aj} == \mu_{Aj} + o_p(1) = \mu + \alpha_j + o_p(1), \qquad (10)$$

and that

$$n^{1/2}(\overline{Y}_A - \mu_A) \rightarrow_d U \qquad (11)$$

where $\overline{Y}_A$ denotes the vector of allelic averages, where U denotes a multivariate normal L-vector, and where $\rightarrow_d$ denotes convergence in distribution. Also,

$$Y'Y/n \rightarrow_p \mu^{(2)} < \infty \qquad (12)$$

and

$$MSE_1 \rightarrow_p \sigma_1^2 > 0 \qquad (13)$$

Proof of Equation (4) now will be provided. Note that:

$$F_1 - F = \frac{MSA_1}{MSE_1} - \frac{MSA}{MSE} = \frac{(MSE)(MSA_1) - (MSE_1)(MSA)}{(MSE_1)(MSE)}.$$

Therefore, if it can be shown

$$MSA - MSA_1 = o_p(1), \qquad (14)$$

$$MSE - MSE_1 = o_p(1), \qquad (15)$$

and

$$MSA_1 \rightarrow_d Q, \qquad (16)$$

for some random variable Q, then

$$F_1 - F = \frac{o_p(1)MSA_1 + o_p(1)MSE_1}{MSE_1(MSE_1 + o_p(1))} = o_p(1)$$

by Equation (13), and the result will be proven. Hence, Equations (14), (15) and (16) need to be demonstrated.

The condition that genotype has no effect on trait implies that $\mu_{jk} \equiv \mu$, and for the remainder of the proof of Equation (4) assume without loss of generality that $\mu = 0$, since all quadratic forms in Equations (14)-(16) are invariant to μ.

To verify Equation (14), it will suffice to show $SSA-SSA_1 = o_p(1)$. Both SSA and SSA_1 are expressed as quadratic forms in $\overline{Y}_A$, then examine the difference of the defining matrices.

Considering SSA first, note that $(A'A)^{-1}A'Y_2 = \overline{Y}_A$. Letting $D_A=A'A=\mathrm{diag}\{n_{Aj}\}$, so $Y_2'A(A'A)^{-1}A'Y_2 = \overline{Y}_A'D_A\overline{Y}_A$. Noting that $J_{1x2n}Y_2=2\Sigma Y_i=2J_{1xn}Y$, and that $J_{1xL}D_A\overline{Y}_A=2\Sigma Y_i$, $Y_2'J_{2nx2n}Y_2/(2n) = \overline{Y}_A'D_AJ_{LxL}D_A\overline{Y}_A/(2n)$. Thus,

$$\begin{aligned} SSA &= \overline{Y}_A'[D_A - D_AJ_{LxL}D_A/(2n)]\overline{Y}_A \\ &= 2(n^{1/2}\overline{Y}_A')\left[\frac{D_A}{2n} - \frac{D_AJ_{Lx1}}{2n}\frac{J_{1xL}D_A}{2n}\right](n^{1/2}\overline{Y}_A) \qquad (17) \\ &= 2(n^{1/2}\overline{Y}_A')A_n(n^{1/2}\overline{Y}_A) \end{aligned}$$

By Equation (9):

$$A_n = P - pp' + o_p(1),$$

where $P = \mathrm{diag}\{p_j\}$ and $p' = (p_1,\ldots, p_L)$.

Now considering $SSA_1 = Y'[D(D'D)^{-1}D'-J_{nxn}/n]Y$, note that $D'Y=D_A\overline{Y}_A$, so that:

$$\begin{aligned} SSA_1 &= \overline{Y}_A'[D_A(D'D)^{-1}D_A - D_AJ_{LxL}D_A/(4n)]\overline{Y}_A \\ &= 2(n^{1/2}\overline{Y}_A')\left[\frac{D_A}{2n}\left(\frac{D'D}{2n}\right)^{-1}\frac{D_A}{2n} - \frac{1}{2}\frac{D_AJ_{Lx1}}{2n}\frac{J_{1xL}D_A}{2n}\right](n^{1/2}\overline{Y}_A) \qquad (18) \\ &= 2\left(n^{1/2}\overline{Y}_A'\right)B_n\left(n^{1/2}\overline{Y}_A\right) \end{aligned}$$

To find the limit of B_n, obtain the limit of $D'D/(2n)$, which may be expressed as:

$$\frac{1}{2n}D'D = \frac{1}{2n}\begin{pmatrix} 4n_{11}+n_{12}+\ldots+n_{1L} & n_{12} & \cdots & n_{1L} \\ n_{21} & n_{21}+4n_{22}+\ldots+n_{2L} & \cdots & n_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ n_{L1} & n_{L2} & \cdots & n_{L1}+n_{L2}+\ldots+4n_{LL} \end{pmatrix}$$

$$= \frac{1}{2n}\mathrm{diag}\{2n_{jj}\} + \frac{1}{2n}\begin{pmatrix} n_{A1} & n_{12} & \cdots & n_{1L} \\ n_{21} & n_{A2} & \cdots & n_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ n_{L1} & n_{L2} & \cdots & n_{AL} \end{pmatrix}$$

$$= P + pp' + o_p(1)$$

under Equations (6)-(9). Since the elements of the inverse of a matrix are continuous functions of the elements of the elements of the original matrix, and since P + pp' is invertible (positive definite in fact), $[(D'D)/(2n)]^{-1} = (P + pp')^{-1} + o_p(1) = P^{-1} - J_{LxL}/2 + o_p(1)$. Thus,

$$\begin{aligned} B_n &= P(P^{-1} - J_{LxL}/2)P - pp'/2 + o_p(1) \\ &= P - pp' + o_p(1) \end{aligned}$$

hence $A_n - B_n = o_p(1)$. Using this result and Equation (11),

$SSA - SSA_1 = 2\left(n^{1/2}\overline{Y}_A'\right)\left[o_p(1)\right]\left(n^{1/2}\overline{Y}_A\right) = o_p(1)$ and Equation (14) is verified.

To show Equation (15), note that

$$\begin{aligned} MSE &= \frac{2n}{2n-L}\left[\frac{Y_2'Y_2}{2n} - \frac{Y_2'A(A'A)^{-1}A'Y_2}{2n}\right] \\ &= (1+o(1))\left[\frac{Y'Y}{n} - \frac{\overline{Y}_A'D_A\overline{Y}_A}{2n}\right] \end{aligned}$$

Also,

$$MSE_1 = (1+o(1))\left[\frac{Y'Y}{n} - \frac{\overline{Y}_A'D_A(D'D)^{-1}D_A\overline{Y}_A}{n}\right]$$

Using Equation (11), as well as convergence results for (D'D)/(2n) and $D_A/(2n)$ given above, it follows that

$$\begin{aligned} MSE\text{-}MSE_1 &= o_p(1) + \overline{Y}_A'\left[\frac{D_A(D'D)^{-1}D_A}{n} - \frac{D_A}{2n}\right]\overline{Y}_A \\ &= o_p(1) + \overline{Y}_A'\left[2(P - pp'/2) - P + o_p(1)\right]\overline{Y}_A \\ &= o_p(1) + \overline{Y}_A'\left[P - pp'\right]\overline{Y}_A \\ &= o_p(1) \end{aligned}$$

since $\overline{Y}_A \rightarrow_p 0$ and Equation (15) is proven.

To show Equation (16), use Equation (18). The result of Equation (16) follows by noting that $n^{1/2}\overline{Y}_A$ converges in distribution and that the elements of B_n converge in probability, and Equation (4) is finally proven.

Proof of Equation (5) now will be provided. Note from Equation (18) that

$$\begin{aligned} SSA_1/(2n) &= \overline{Y}_A'B_n\overline{Y}_A \\ &\rightarrow_p \mu_A'(P - pp')\mu_A = V\alpha . \end{aligned}$$

From Equation (17)

$$SSA/(2n) = \overline{Y}'_A A_n \overline{Y}_A \rightarrow_p V_\alpha$$

Hence, $SSA_1/SSA \rightarrow_p 1$.

Now consider

$$\frac{MSE}{MSE_1} = \frac{MSE - MSE_1}{MSE_1} + 1$$

From Equations (10), (13) and (7),

$$\frac{MSE - MSE_1}{MSE_1} \rightarrow_p \frac{V_\alpha}{\sigma_1^2},$$

implying $MSE/MSE_1 = V_\alpha/\sigma_1^2 + 1 + o_p(1)$, and the result is proven.

Excoffier and Slatkin (1995) Molecular Biology and Evolution 12, 921-7 suggested that the variance estimates for haplotype frequencies can be obtained by inserting in the final frequency estimates into the information matrix and inverting it. Thus, a score test can be constructed for testing the association of the presence of disease with frequencies of haplotypes. Another approach, as have been used in Zhao et al. (2000) Human Heredity 50, 133-9, is to estimate haplotype frequencies separately in cases, controls, and in the pooled sample, compute three corresponding sample log-likelihoods (L1, L2, L3), and then form a likelihood ratio (LR) test statistic as 2(L1+L2-L3). This statistic is asymptotically chi-squared with the degrees of freedom given by the number of haplotypes minus one. Having parental information allows for more efficient haplotype inference. Clayton (1999) American Journal of Human Genetics 65, 1170-7 and accompanying program "Transmit" is relevant for two generation pedigrees.

The contents of all of the above publications (as well as those mentioned in the Examples) concerning algorithms are incorporated herein by reference.

The haplotype analysis is generally performed over a sliding scan window, such as for polymorphisms from a whole genome scan, chromosome scan or a chromosomal region scan. The haplotype scan window is generally at least 1 kb or 10 kb, and is preferably 10 kb to 100 kb, 10 kb to 500 kb, 10 kb to 1000kb, 10 kb to 10,000 kb, 500 kb to 10,000 kb or larger. Typically the scan window comprises at

least 2, 3, 4, 5, 10, 20 or more polymorphic regions (e.g. SNPs). Generally at least 2, 3, 5 or more different scan windows are used.

The length of the scan window might be analysis-specific, and analysis with different window lengths can be used to fully explore the data. Generally, marker in high pair-wise linkage disequilibrium would provide similar information and one of such markers might be excluded form consideration. The pair-wise disequilibrium might also provide some insight on the optimal length of the window, as there might be no reason to extend it over distances where linkage disequilibrium substantially drops. Windows containing larger numbers of markers might provide more precise information on the association with the response, but the estimation precision and high statistical power will require larger sample sizes.

The statistical analysis which detects the correlation between the phenotype and deduced haplotypes determines whether an individual who has the haplotype is more likely to have the relevant phenotype. Generally this is done by determining whether there is a difference in the phenotype amongst people with and without the haplotype. In a case-control study the frequency of the haplotype in the cases is compared to the frequency of the haplotype in the controls. Typically a likelihood ratio test is used and the p-value for the association is obtained from the chi-square distribution.

The invention is illustrated by reference to the following Example:

Examples

Materials and Methods

DNA samples

Blood was collected from 81 GlaxoSmithKline employees (Caucasians) and 154 males (Caucasian) from the United States under informed consent. DNA was extracted by PPGX (Research Triangle Park, North Carolina) using the Puregene DNA isolation kit (Gentra Systems Inc, supplied by Flowgen Instruments Ltd, Lichfield, UK) or by Whatman Bioscience (Cambridge, UK). All DNA samples were anonymised.

Genotyping

Five SNPs from the *N-Acetyltransferase* 2 (*NAT2)* gene on chromosome 8 (table 1) were genotyped for each of the 81 GlaxoSmithKline employees using PCR and direct sequencing. An 850-bp fragment of the *NAT2* gene was amplified using

primers F1 and R1 and subsequently sequenced using the initial PCR primers and two additional nested primers (F2 and R2) on an ABI 377 Sequencer (PE Applied Biosystems, Foster City, USA). The sequences of the primers were as follows:

F1 (forward PCR primer): 5'-CTATAAGAACTCTAGGAACAAATTGGAC-3'

R1 (reverse PCR primer): 5'-AAGGGTTTATTTTGTTCCTTATTCTAAAT-3'

F2 (nested forward primer): 5'-CACCTTCTCCTGCAGGTGACCA-3'

R2 (nested reverse primer): 5'-TGTCAAGCAGAAAATGCAAGGC-3'

Sequencher (Genecodes Corporation, Ann Arbor, USA) was used to analyse the sequences in order to generate genotype results for each of the 5 polymorphic sites.

Five SNPs over a region of 140-kb on the X chromosome were identified by PCR and direct sequencing of DNA samples from 11 female individuals (Coriell Cell Repositories, New Jersey, USA). Oligo ligation assays (OLA) were used to generate genotype calls for the 154 males. Table 2 shows the sequences of the primers and probes used in the PCR and OLA assays. OLA-PCR was performed at 94°C for 2 minutes, then at 94°C for 30 seconds, 50°C for 30 seconds, 72°C for 1 minute for 40 cycles, and finally at 72°C for 5 minutes using an MJ thermal cycler (MJ Research INC, Watertown, Massachusetts, USA).

This was followed by a heat-kill step of 99°C for 30 minutes to inactivate the remaining viable Taq polymerase. The ligation reaction was run at 94°C for 20 seconds and 50°C for 1 minute for 30 cycles on an MJ thermal cycler. Ten µl reaction mix contained 3 µl of the lyophilised PCR products, 10 units of Taq DNA ligase (thermo-stable), 45 nM of each of the three probes, and 1X ligase buffer (20mM Tris-HCl, 25mM potassium acetate, 10mM magnesium acetate, 10mM DTT, 1mM NAD, and 0.1% Triton X-100). Subsequently, 2 µl of the ligation products were mixed with 18 µl of a size standard, and 2 µl of this were loaded onto an ABI 3700 sequencer (PE Applied Biosystems, Foster City, USA) to separate the different alleles. The data was analysed using Genotyper NT (PE Applied Biosystems, Foster City, USA) to generate genotype calls.

Molecular determination of the haplotypes

The 850-bp PCR fragment of the *NAT2* gene was cloned into a TA cloning vector (Invitrogen, Groningen, the Netherlands). Between six and twelve subclones

from each of the 81 individuals were sequenced. These sequence data were analysed using Lasergene (DNASTAR Inc., Madison, USA) to resolve the haplotypes for both chromosomes of each individual. The haplotypes from the 5 SNPs on chromosome X were assigned directly according to the genotype data, as each individual male has only one X chromosome.

Computational estimation of the haplotypes

For the chromosome X region, artificial diploid genotypes were constructed by combining random pairs of males. The haplotypes for each diploid for both genetic regions were assigned using the subtraction algorithm (Clark (1990) Mol. Biol. Evol. 7, 111-22). Briefly, haplotypes for individuals who were either complete homozygotes or single - site heterozygotes were assigned initially and a preliminary list of haplotypes present in the samples was recorded. Then other individuals who carried a copy of the previously recognised haplotypes were identified. Each time a resolved haplotype was identified as one of the possible alleles in an ambiguous individual, the homologue allele was considered to be a recognisable haplotype and added to the haplotype list. This exercise was repeated until the phase information for all individuals was either resolved or identified as unresolved. Depending on the order in which the genotypes are entered, the algorithm may produce a different set of haplotypes. The haplotype frequencies were calculated by gene counting using individuals with resolved haplotype phases.

The sample haplotype frequencies and individual conditional haplotype probabilities for both genomic regions were also estimated using the EM algorithm with multiple restarts. All haplotype pairs that can yield an unphased genotype pattern were enumerated. The probability for each of the haplotype configurations was calculated using the estimated population haplotype frequencies. The haplotype phase was considered to be resolved if the probability of a haplotype pair was greater than 99%. For example, suppose one haplotype pair that generates the unphased pattern is i/j, where i and j represent two of the haplotypes with $p(i)$ and $p(j)$ frequencies as estimated by the EM algorithm. By the Bayes rule, the conditional probability that the unphased genotype G_{ij} has the haplotype pair i/j is

$$\Pr\left(p(i), p(j) \mid G_{ij}\right) = \frac{p(i)p(j)}{\sum_{x,y} p(x)p(y)}$$

where x and y indicate a haplotype pair that can yield the same unphased genotype, and the sum is taken over all such pairs including i and j. If the conditional probability is less than 99%, the phase of that genotype pattern is considered unresolved.

Measures of estimation accuracy

The two measures, I_F and I_H, introduced by Excoffier and Slatkin were used to estimate the effectiveness of computational algorithms when predicting haplotype frequencies (Excoffier and Slatkin (1995) Mol. Biol. Evol. 12, 921-7). I_F (the similarity index) describes how close the estimated haplotype frequencies are to the actual frequencies, and is defined as the proportion of haplotype frequency in common between estimated and true frequencies.

$$I_F = \sum_{k=1}^{h} \min(p_{ek}, p_{tk}) = 1 - \frac{1}{2}\sum_{k=1}^{h} |p_{ek} - p_{tk}|$$

Where h is the number of haplotypes in the data set, p_{ek} and p_{tk} are the estimated and true (experimentally determined in this case) haplotype frequencies for the k haplotype, k=1....h. I_F varies between 0 and 1 (a value of 1 is achieved when the actual and estimated frequencies are identical). I_H compares the number of different haplotypes seen experimentally with the number of different haplotypes identified by the computer programs. A haplotype is defined as being detected if it has an estimated frequency of at least 1/(2n) in a population of n individuals (Excoffier and Slatkin, supra).

$$I_H = \frac{2(m_{true} - m_{missed})}{m_{true} + m_{est}}$$

Where m_{true} is the number of haplotypes determined experimentally, m_{est} is the number of estimated haplotypes with frequency above the threshold, and m_{missed} is the number of haplotypes identified experimentally but not computationally. The value of I_H can vary between one (when the computational identified haplotypes are exactly the same as those determined experimentally) to zero (when none of the true haplotypes are identified computationally).

The mean squared error (MSE) described by Fallin and Schork was also used to measure the accuracy of computational algorithms in haplotype frequency estimation (Fallin and Schork (2000) Am. J. Hum. Genet. 67, 947-59). The MSE measure incorporates all the k haplotype frequencies and thus reflects the overall

difference in haplotype frequencies between estimated and true values for a particular data set.

$$MSE = \sum_{k=1}^{h} (p_{ek} - p_{tk})^2 / h$$

Where h, p_{ek} and p_{tk} are defined as above.

Pair-wise Linkage Disequilibrium between SNP markers

LD was measured using the standardised D' first proposed by Lewontin (Lewontin (1964) Genetics 49, 49-67). D' is the LD relative to its maximum value for a given set of allelic frequencies for the pair of sites. It is calculated by dividing the raw D value by the absolute maximal value possible. In that sense, D' is a normalised value of LD.

Results

Molecular determination of genotypes and haplotypes for the *NAT2* locus

Figure 1 shows the distribution of the 5 SNPs utilised in this study over the *NAT2* locus. The 850-bp fragment of the *NAT2* gene was amplified using primers F1 and R1 and the genotypes of each individual for the 5 SNPs were determined by direct sequencing of the PCR products in both directions. The minor allele frequencies of the 5 polymorphisms determined in the 81 individuals ranged from 0.25 to 0.49 (table 1), which were similar to those reported for Caucasians (Agundez et al (1996) Pharmacogenetics 6, 423-8). The genotype distribution for each SNP did not deviate significantly from Hardy-Weinberg equilibrium.

To determine the haplotypes molecularly, the 850-bp PCR fragment was cloned into a TA cloning vector and between six and twelve subclones from each individual were analysed by PCR and sequencing. In the absence of recombination, recurrent mutation and back mutation, the maximum number of haplotypes for a locus with 5 biallelic variable sites is 6 (i.e. n+1), with n being the number of SNP sites. On the other hand, if there is random association between polymorphic sites, the maximum number of potential haplotypes for a locus with 5 SNPs is 32 (2^5). Analysis of the 162 alleles in the *NAT2* locus revealed 7 haplotypes suggesting strong LD in this small chromosomal region (table 3). Indeed, there were maximal or nearly maximal D' values among all SNP pairs indicating that there was complete or near complete linkage disequilibrium and that recombination was rare over such a short

physical distance in the *NAT2* gene locus (table 4, figure 1). The 5 SNPs generated ten SNP pairs (table 4). Each of the eight SNP pairs created only three haplotypes. The remaining two SNP pairs created all four possible haplotypes with three haplotypes accounting for 98-99% of the alleles.

Computational estimation of haplotypes for the *NAT2* locus

We inferred haplotypes from the genotyping results for the 81 individuals using the subtraction algorithm (Clark (1990), supra). Thirty-one individuals were either homozygous for all SNP sites or heterozygous at only one SNP site; thus their haplotypes could be assigned directly. Six haplotypes were identified in these 31 individuals (table 3, H1-H6). Eight, eighteen, three and twenty-one individuals were heterozygous at two, three, four and all five SNP sites respectively. Using the subtraction method, we resolved the haplotype phases for 64 individuals (79%). There was 100% concordance between experimentally determined haplotype phases and those predicted computationally for the individuals.

The remaining 17 individuals were heterozygous at the same three SNP sites and each had two possible haplotype configurations. The haplotype frequencies were calculated from the 64 phase-resolved individuals (table 3). The similarity index (I_F) value was 0.91, which was close to its maximal value, suggesting that the subtraction method was effective in estimating haplotype frequencies for this region. The overall error (MSE value) between estimated and true sample frequencies was 1-2 orders of magnitude greater than that reported by Fallin and Schork using the EM algorithm (Fallin and Schork (2000), supra), probably because of the subtraction method used a reduced number of individuals in haplotype frequency estimation (table 3).

We also estimated the haplotype frequencies using the EM algorithm with 100 restarts to minimize chances of local convergence. The algorithm predicted a total of 7 haplotypes with 3 main haplotypes (H1-H3) representing 93% of all alleles (table 3). Comparison of the haplotype frequencies determined molecularly and that estimated using the EM algorithm showed very high concordance (table 3). The I_F value was 0.999 and the MSE value was 4 orders of magnitude smaller than that obtained using the subtraction method. We calculated the conditional probability for each of the haplotype configurations for each individual from the estimated haplotype frequencies. We considered haplotype phases to be resolved if the probability of a haplotype pair was greater than 99%. The EM algorithm predicted haplotypes phases

for all of the 81 individuals. There was 100% agreement between the predicted haplotypes and that determined experimentally, indicating pronounced accuracy for haplotype assignment for this region. Our results suggested that the EM algorithm performed better than the subtraction algorithm in both estimating haplotype frequencies and predicting individual haplotype phases for the *NAT2* locus.

Molecular determination of genotypes and haplotypes for the X chromosome locus

Figure 2 shows the distribution of the 5 SNPs over a region of 140-kb on chromosome X. The 5 SNPs for the 154 males were genotyped using OLA. The minor allele frequencies of the 5 SNPs ranged from 0.06 to 0.40 (figure 2). Pair-wise linkage disequilibrium was measured using D' (table 5).

The haplotypes of the 5 markers for the 154 males were assigned directly according to the genotype data, as each individual male has only one X chromosome. The five polymorphisms established 21 out of the 32 (2^5) potential haplotypes (table 6). Six of the haplotypes (h16-h21) were observed only once and four haplotypes (h12-h15) were seen only twice. These ten rare haplotypes (h12-21) represented 9% of all the 154 alleles. Six haplotypes (h1-h6) had allele frequencies above 5%, representing 75% of the 154 alleles. Of these six haplotypes, four had allele frequencies above 10% (h1-h4), accounting for 57% of all alleles.

Computational prediction of haplotypes for the chromosome X locus

To evaluate the effectiveness and accuracy of the computational methods in predicting haplotype phases and estimating haplotype frequencies over a relatively large genetic region (140-kb), we artificially created genotype data for 77 diploids by combining random pairs of males. According to the subtraction algorithm (Clark (1990), supra), we resolved the haplotype phases for a total of 43 diploids (56%), including 38 diploids which were either complete homozygotes or single-site heterozygotes and 5 diploids which were heterozygous at multiple SNP sites (table 7). There was more than one possible haplotype configuration for each of the remaining 34 diploids (44%), and the haplotype phases of these diploids remained unresolved (table 7). To evaluate the accuracy of the subtraction algorithm in predicting haplotype phases, we made direct comparisons between the computationally predicted haplotypes and that experimentally determined. For diploids that were either complete

homozygotes or single-site heterozygotes, there was 100% match between the two sets of data. However, the computationally assigned haplotype configurations were in agreement with that experimentally determined for only 3 out of the 5 multiple-site heterozygotes (table 7). Our data suggested that the subtraction method was neither effective nor accurate in predicting haplotype phases for diploids that were heterozygous at multiple SNP sites in genomic regions where pronounced LD was not maintained.

The haplotype frequencies were estimated using the 43 phase-resolved diploids by the subtraction method (table 6). The combined haplotype frequency for the six haplotypes (h1-h6) that had true allele frequencies greater than 5% was 0.81, which was higher than that determined molecularly (0.75). The I_F value of the subtraction method was lower for this region than that for the *NAT2* region (table 3, 6).

We also estimated the haplotype frequencies for the 77 artificially generated diploids using the EM algorithm with 100 restarts. A total of 26 haplotypes were predicted, including the 21 molecularly determined haplotypes and 5 additional rare haplotypes which accounted for 0.1% of all the alleles (table 6, data not shown). Sixteen haplotypes predicted by the EM algorithm had allele frequencies greater than 0.5%. For the 10 haplotypes (h12-21) observed only once or twice molecularly, zero to 1.5 alleles were predicted computationally, accounting for 4% of all the alleles. For the six haplotypes (h1-h6) that had true allele frequencies above 5%, haplotypes predicted computationally represented 76% of the 154 alleles, which is similar to that determined molecularly (75%). The EM algorithm performed marginally better than the subtraction method in estimating haplotype frequencies for this locus (table 6). The reduced I_F value and increased MSE value for this locus in comparison with that observed for the *NAT 2* locus suggested that the estimation error for overall haplotype frequencies was increased with decreased LD between SNP sites using the EM algorithm (table 3, table 6).

We calculated the conditional probability of individual haplotype configuration for each diploid from the estimated haplotype frequencies using the EM algorithm. The haplotype phases for 49 out of the 77 diploids (64%) were considered to be resolved. More than one possible haplotype configuration was present for the remaining 28 diploids (36%), and the haplotype phases for these diploids remained unresolved (table 7). The EM algorithm predicted haplotypes were in agreement with

those experimentally determined for all of the 38 diploids, which were either complete homozygotes or single-site heterozygotes. For the 11 diploids that were multiple-site heterozygotes, the haplotype phases assigned by the EM algorithm were in agreement with those experimentally determined for only 5 diploids. The overall accuracy for predicting haplotype phases using the EM algorithm was 88% for all of the diploids in this region. Thus, the EM algorithm also performed poorly in predicting individual haplotype phases for genetic regions with low LD.

To make a more comprehensive comparison between the haplotype subtraction method and the EM method in haplotype frequency estimation, we performed simple computer simulation to assess the accuracy of both algorithms when there is uniformity of haplotype frequencies. The amount of heterozygosity, and therefore the number of ambiguous haplotypes, was increased by equalising haplotype frequencies, thereby presenting more of a challenge to the computational algorithms. We took the empirical pool of haplotypes in table 6 and assumed equal haplotype frequencies for all haplotypes (1/21). Random samples of 77 individuals were taken from populations with these frequencies, assuming random union of haplotypes. Sample haplotypes were estimated from genotypes through the EM algorithm with 100 random re-starts (table 8). In comparison with the results presented in table 6, there was a decrease in I_F value for the simulated data set, indicating that there was increased estimation error for haplotype frequencies using the EM algorithm when the haplotype frequencies reached uniformity. Using the subtraction method, the haplotype phases were resolved for only 21/77 (27%) diploids, and the haplotype frequencies were calculated using these phase-resolved diploids (table 8). This exercise suggested that there was increased estimation error for haplotype frequencies with increased ambiguity using both computational methods.

Discussion

We evaluated the effectiveness and accuracy of two computational algorithms in estimating haplotype frequencies and in predicting haplotype phases using molecular data. We experimentally determined individual genotypes and haplotypes for 5 SNPs over a region of 850-bp and 5 SNPs over a region of 140-kb. The subtraction method and an EM algorithm based computational method were applied to estimate the haplotype frequencies and predicting individual haplotypes from the genotype data (Clark 1990, supra; Chiano and Clayton (1998) Am. Hum. Genet. 62,

55-60). Direct comparison of the computationally predicted haplotype frequencies and individual haplotypes with that determined experimentally allowed us to evaluate and compare the effectiveness and accuracy of both algorithms. To our knowledge, this is the first study to evaluate and compare computational algorithms in both haplotype frequency estimation and individual haplotype assignment using experimental data.

We have found that both computational methods performed well in overall haplotype frequency estimation for genetic regions with high LD (table 3). The accuracy of computational methods were decreased with decreased LD and increased ambiguity (table 3, 6, 8). The EM algorithm gave better overall estimates in haplotype frequencies than the subtraction method for both genomic regions (table 3, 6). This may reflect the fact that the EM algorithm included all individuals from the samples in the haplotype frequency estimation, whereas the subtraction method only used the phase-resolved individuals. The proportion of phase-resolved individuals was reduced with decreased LD between SNP sites (table 9). Both algorithms gave better estimates in haplotype frequency for the genomic region with pronounced LD (the *NAT2* locus) than that for the region where substantial LD is not maintained (chromosome X).

This observation is in agreement with that reported recently in a simulation study to assess the accuracy of the EM algorithm in haplotype frequency estimation (Fallin and Schork 2000, supra). Fallin and Schork demonstrated that the EM algorithm performed very well under a wide range of population and data set scenarios. We have shown that haplotype frequencies can be estimated from genotype data computationally without additional laboratory cost, and the estimation error was increased with decreased LD.

We also evaluated and compared the effectiveness and accuracy of both computational algorithms in predicting haplotype phases for individuals. Both algorithms predicted individual haplotypes effectively and accurately for the *NAT2* region, where there was near complete LD between SNP sites (table 3, 9). The EM algorithm gave a better estimate than the subtraction method. Such effectiveness and accuracy in haplotype prediction were reduced when marked LD was not maintained, as demonstrated over the chromosome X locus (table 6,7,9). The subtraction and EM algorithms resolved haplotype phases for 56% and 64% of the diploids respectively and the accuracy for the two algorithms was 95% and 88% respectively (table 9). Our results indicated that the computational algorithms could provide effective and accurate prediction for haplotype phases in genetic regions with pronounced LD but

not in regions where marked LD is not maintained. It needs to be highlighted that the degree of inherent phase ambiguity for multiple-site heterozygotes is increased with decreased LD between markers (Hoh and Hodge 2000). The performance of computational algorithms in predicting haplotype phases should be interpreted in the light of this inherent ambiguity.

Our observations may have potential implications for genome - wide association studies. Sequential examination of individual SNPs in an attempt to identify disease-susceptibility genes is fraught with problems of interpretation. Firstly, the number of analyses performed will be enormous for the 100,000-500,000 SNPs which will be used in such studies, making it necessary to correct the statistical result and to ensure the authenticity of the signal from each SNP.

Secondly, in contrast with monogenic diseases where the causative single nucleotide changes may have unambiguous phenotypes, the contributions of genetic variations in the underlying network of interactions that are responsible for the phenotypes of complex diseases are much more complicated. The pattern of genotype-phenotype association might be more complex than initially envisaged. Indeed, the very fact that a large number of SNPs have been identified within coding and regulatory regions of a specific candidate gene, raises the possibility that several of them within a gene might be functional.

Thirdly, current knowledge about the actual distribution of LD across the human genome is limited. If complete or near complete association between several polymorphisms within a gene is present, haplotypic combinations of the polymorphic sites may play a significant role in the functionality of the gene. Therefore, it is necessary to explore multiple alternative analytical approaches to identify disease genes in association studies. Incorporating haplotype analyses will ensure additional use of valuable information in association studies and provide additional evidence about the strength and nature of the associations. Once a collection of SNPs has been discovered and genotyped over a gene locus, a chromosomal region, or even the entire genome, they can be organised sequentially into haplotypes. This will allow sequential haplotype scans for two, three, four, five or more SNPs on fragments of DNA ranging from 10-150 kb in association studies. Sequential haplotype scanning may be able to provide a richly detailed view of specific genomic fragments and reveal the inter-relationships between SNPs surrounding the regions, thus offering an additional method for identifying genomic fragments that harbour the variants causing

the phenotype. If the haplotype information is derived from genotype data using computational methods, it needs to be noted that the accuracy of such haplotype information is decreased with decreased LD and increased ambiguity between markers.

Our observations also have potential implications for linkage studies using SNPs. Currently, approximately 400 microsatellite markers are used for genome - wide linkage scans to localise regions harbouring disease genes, with an average genetic distance being approximately 10 cM. Due to the remarkable advances in the technologies for SNP identification and genotyping, it is proposed that it may be more efficient to use 1500-2000 SNP markers to replace microsatellite markers in a typical genome scan. The combined polymorphic information content from several highly informative SNPs within a region may be equivalent to the polymorphic information content from a single multi-allelic microsatellite marker. For a locus with n biallelic variable sites, the maximum number of haplotypes is $n+1$ in the absence of recombination, repeated or back mutations; whereas the potential number of haplotypes could reach 2^n if there is linkage equilibrium between polymorphic sites. We observed 7 and 21 haplotypes for the 5 SNPs over the *NAT2* locus and the X chromosome locus respectively in the populations studied, reflecting the different magnitudes of LD operating over the 850-bp and the 140-kb regions. The haplotype heterozygosity (H) in the X chromosome region (H=0.82, based on the artificial diploid data) is higher than that in the *NAT2* locus (0.69), albeit the opposite is true for average H of individual SNPs over the two regions (0.31 and 0.48 respectively).

This is consistent with the strong negative correlation between the mean pair-wise LD and haplotype heterozygosity, supporting the concept that the stronger the non-random association between SNPs, the lower the information added by each SNP to a set of other SNPs. On the other hand, some information will be lost due to ambiguous haplotyping of multiple SNPs, if linkage equilibrium is reached between SNP sites. A fine balance of LD between multiple SNP sites is therefore required to get the maximum information within a genomic region, and the knowledge of the extent and magnitude of LD between SNPs will be invaluable in the selection of the SNP set for linkage analysis.

This study indicated that computational methods provide effective prediction of haplotype frequencies using genotype data from unrelated individuals for genomic regions with LD. Computational algorithms give effective and accurate prediction for

haplotype phases in regions with high values of LD between markers and small probability of recombination events. The EM algorithm is a better computational method than the subtraction method both in estimating haplotype frequencies and in predicting haplotype phases. Our observation provides alternative statistical approaches in association studies and linkage studies using SNPs.

Table 1 Allele frequencies of the 5 SNPs of the *NAT2* gene in 81 individuals

SNP ID	Nucleotide position (U53473)	Base change	Minor allele frequencies
SNP1	282	C-T	0.27 (T)
SNP2	341	T-C	0.49 (C)
SNP3	481	C-T	0.48 (T)
SNP4	590	G-A	0.25 (A)
SNP5	803	A-G	0.46 (G)

Table 2 OLA primer and probe sequences for the 5 SNPs on the X chromosome

	Sequences
SNP6 (T/G)	
F PCR primer	TTTTGGTGTTGCAGTATTGACAG
R PCR primer	TCTTGGGAAGCATAGGTCTCTTG
Allele 1 probe	GCTGTCAGAACAGGAATT (FAM)
Allele 2 probe	ACAGCTGTCAGAACAGGAATG (FAM)
Common probe	TCCAAACTGCTCTAGCTGAAGACAG
SNP7 (G/T)	
F PCR primer	CCACAAATCTTTGCTGTGATGAG
R PCR primer	ACCCCATGCTAGACATGCTATTC
Allele 1 probe	AATGAGTGGTCCGGGAAG (HEX)
Allele 2 probe	CAAAATGAGTGGTCCGGGAAT (HEX)
Common probe	CCCTTGCTATAGACGGGAGAATGCTACAGTCTC
SNP8 (A/G)	
F PCR primer	GAGCTGGAAAGCACCAGAACATG
R PCR primer	GAGGCGATCTCCAGCCTCC
Allele 1 probe	TCCTTTTCCCAAACCAGA (FAM)
Allele 2 probe	TCATCCTTTTCCCAAACCAGG (FAM)
Common probe	GCTCTATATGTTCAAGGAAATGCAGCGGTATGTGTGCCT
SNP9 (C/G)	
F PCR primer	AGACACGAAGGAGTGCATTCTG
R PCR primer	TCTAGCCCAAACCTCTTTTGAAG
Allele 1 probe	TTACAAAGTCAACTCACC (HEX)
Allele 2 probe	TTTTTACAAAGTCAACTCACG (HEX)
Common probe	CGTTAGCCACCACTGAGATCAAGAGC
SNP10 (T/C)	
F PCR primer	CCACATAGATGCTTCCAGCAGC
R PCR primer	GTTCAGTTTTGCCTGACGATC
Allele 1 probe	AATGCTACAGAGAAGCTT (FAM)
Allele 2 probe	AAGAATGCTACAGAGAAGCTC (FAM)
Common probe	AAGTAGTGAACATAGTGGGGAGCTTGAGTCAC

Table 3 Haplotype frequencies determined by molecular and computational methods for the *NAT2* locus (n=81 individuals)

Haplotype	Haplotype[a]	Experimental determined frequency	Estimated frequency (subtraction algorithm)[b]	Estimated frequency (EM algorithm)[c]
H1	12212	0.444	0.430	0.444
H2	11111	0.235	0.164	0.234
H3	21121	0.247	0.313	0.247
H4	21111	0.025	0.031	0.025
H5	12211	0.031	0.039	0.031
H6	12112	0.012	0.016	0.012
H7	11112	0.006	0.008	0.005
I_F			0.914	0.999
I_H			1	0.923
MSE			1.4E-03	2.9E-07

[a] Allele 1 refers to the major allele and allele 2 refers to the minor allele for all the 5 SNPs in this locus.

[b] The haplotype phases were resolved for 64/81 individuals according to the subtraction method (Clark 1990). The haplotype frequencies were calculated from the 64 phase-resolved individuals. The remaining 17 phase unresolved individuals were triple heterozygotes with the genotype distribution being 11,12,12,11,12.

[c] The E-M algorithm with 100 restarts.

Table 4 Linkage disequilibrium (absolute D' value) between SNP markers in the *NAT2* locus

	SNP2	SNP3	SNP4	SNP5
SNP1	1	1	1	1
SNP2		1	1	0.97
SNP3			1	0.92
SNP4				1

Table 5 Linkage disequilibrium (absolute D') between SNP markers over the chromosome X locus

D'	SNP7	SNP8	SNP9	SNP10
SNP6	0.27	0.40	0.59	0.17
SNP7		0.23	0.1	0.40
SNP8			0.23	0.21
SNP9				0.71

Table 6 Haplotype frequencies determined by molecular and computational methods for the chromosome X region

Haplotype	Haplotype[a]	Actual frequency	Estimated frequency (subtraction algorithm)[b]	Estimated frequency (EM algorithm)[c]
h1	12212	0.169	0.186	0.176
h2	12211	0.162	0.221	0.172
h3	11211	0.123	0.128	0.122
h4	12112	0.117	0.081	0.095
h5	12221	0.097	0.081	0.117
h6	12111	0.084	0.116	0.080
h7	11221	0.045	0.035	0.039
h8	11112	0.039	0.047	0.053
h9	12121	0.032	0.012	0.033
h10	11121	0.019	0.023	0.023
h11	22112	0.019	0.035	0.045
h12	11212	0.013	0.012	0.010
h13	11222	0.013	0.000	0.007
h14	12222	0.013	0.012	0.009
h15	22212	0.013	0.000	2.76E-6
h16	11111	0.006	0.000	4.23E-10
h17	21111	0.006	0.012	0.007
h18	21211	0.006	0.000	0.011
h19	22111	0.006	0.000	4.14E-6
h20	22121	0.006	0.000	7.19E-10
h21	22211	0.006	0.000	1.75E-15
I_F			0.856	0.914
I_H			0.800	0.865
MSE			3.68E-04	1.13E-04

[a] For SNP 6, SNP9, and SNP10, allele 1 refers to the major allele and allele 2 refers to the minor allele. For SNP7 and SNP 8, allele 1 refers to the minor allele and allele 2 refers to the major allele.

[b] According to the subtraction method (Clark 1990). The haplotype frequencies in this column were calculated from 43/77 diploids. The phases of the remaining 34 diploids (19 double heterozygotes, 12 triple heterozygotes and 3 quadruple heterozygotes) remained ambiguous, i.e. there were more than one possible haplotype configurations.

[c] The E-M algorithm with 100 restarts.

Table 7 Haplotype assignments for individuals who were heterozygous at multiple SNP sites by molecular and computational methods for the chromosome X region

Genotypes					n	Haplotypes (subtraction)	Haplotypes (EM)	Haplotypes (experimental)
SNP6	SNP 7	SNP 8	SNP 9	SNP 10				
11	11	12	11	12	1	11112/11211	11112/11211	11112/11211
11	12	22	11	12	2	unresolved	unresolved	12212/11211
11	12	22	12	11	6	unresolved	unresolved	11211/12221[a]
11	22	22	12	12	2	unresolved	unresolved	12221/12212
11	12	12	11	11	1	11211/12111	11211/12111	12211/11111[h]
11	22	12	12	11	1	unresolved	unresolved	12221/12111
11	22	11	12	12	1	unresolved	12121/12112	12121/12112
12	22	11	11	12	2	12111/22112	12111/22112	12111/22112[b,h]
11	22	12	11	12	3	unresolved	unresolved	12112/12211[c]
11	12	12	11	22	1	unresolved	unresolved	12212/11112
11	12	12	22	11	1	unresolved	unresolved	12121/11221
11	12	22	12	22	1	unresolved	unresolved	11222/12212
12	22	12	11	22	1	unresolved	12212/22112	12112/22212[h]
12	11	12	11	11	1	11211/21111	11211/21111	11211/21111
11	12	12	11	12	3	unresolved	unresolved	11211/12112[d]
11	22	12	12	12	4	unresolved	unresolved	12112/12221[e]
12	11	12	11	12	1	unresolved	unresolved	11112/21211
12	22	11	12	12	1	unresolved	12121/22112	12112/22121[h]
12	22	12	11	12	2	unresolved	12211/22112	12211/22112[f,h]
12	12	22	11	12	1	unresolved	12212/21211	22212/11211[h]
11	12	12	12	12	3	unresolved	unresolved	11221/12112[g]

[a] Four individuals had haplotype composition 11211/12221, and two individuals had haplotype composition 12211/11221.

[b] One individual had haplotype composition 12111/22112, and one individual had haplotype composition 22111/12112.

[c] Two individuals had haplotype composition 12112/12211, and one individual had haplotype composition 12111/12212.

[d] Two individuals had haplotype composition 11211/12112, and one individual had haplotype composition 11212/12111.

[e] One individual had haplotype composition 12112/12221, two individuals had haplotype composition 12121/12212, and one individual had haplotype composition 12222/12111.

[f] One individual had haplotype composition 12211/22112, and one individual had haplotype composition 12112/22211.

[g] One individual had haplotype composition 11221/12112, one individual had haplotype composition 12212/11121, and one individual had haplotype composition 11222/12111.

[h] There were discrepancies between molecularly determined haplotypes and computationally predicted haplotypes.

Table 8 Haplotype frequency estimation using computational algorithms for the chromosome X region assuming uniform population haplotype frequencies from simulated data (N=77)

Haplotype	Sample frequency	Estimated frequency (Subtraction method)[a]	Estimated frequency (EM algorithm)[b]
11111	0.039	0.024	0.012
11112	0.045	0.024	0.042
11121	0.039	0.024	0.016
11211	0.097	0.071	0.120
11212	0.045	0.071	0.036
11221	0.026	0.024	0.030
11222	0.039	0.048	0.040
12111	0.026	0.024	0.009
12112	0.058	0.095	0.078
12121	0.084	0.071	0.116
12211	0.026	0.071	0.063
12212	0.045	0.071	0.038
12221	0.058	0.048	0.036
12222	0.032	0.024	0.015
21111	0.032	0.000	0.092
21211	0.045	0.000	0.0001
22111	0.064	0.048	0.021
22112	0.064	0.095	0.049
22121	0.039	0.048	0.048
22211	0.039	0.048	0.041
22212	0.058	0.071	0.072
I_F		0.795	0.785
I_H		0.950	0.976
MSE		5.40E-04	6.56E-04

[a] The haplotypes were assigned according the subtraction method (Clark 1990). The haplotype frequencies in this column were calculated from 21/77 phase-resolved diploids.

[b] The E-M algorithm with 100 restarts.

Table 9 Comparison of computational methods in predicting haplotype phases

Method	*NAT2* (n=81)		Chromosome X (n=77)	
	Phase-resolved individuals (%)	Accuracy[a]	Phase-resolved diploids (%)	Accuracy[a]
Subtraction	64 (79%)	100%	43 (56%)	95%
EM	81 (100%)	100%	49 (64%)	88%

[a] We calculated overall accuracy from phase-resolved individuals.

CLAIMS

1. Method of performing an association study comprising:

(a) obtaining information about (i) genetic polymorphisms and (ii) phenotypes which are present in a sample population

(b) performing a haplotype analysis on the genetic polymorphism information to deduce the haplotypes present in the population, and

(c) performing a statistical analysis to detect a correlation between the phenotypes and the deduced haplotypes,

thereby determining whether there is an association between the polymorphisms and the phenotypes.

2. Method according to claim 1 wherein the information about the genetic polymorphisms does not indicate the phase of the polymorphisms.

3. Method according to claim 1 or 2 wherein the information about the genetic polymorphisms indicates the polymorphisms present across a region of at least 10 kb of the genome of the individuals in the study.

4. Method according to any one of the preceding claims wherein information about the genetic polymorphisms comprises an indication of the polymorphisms present in a region of at least every chromosome of individuals in the study.

5. Method according to any one of the preceding claims wherein the association study is a case-control study in which the frequency of the haplotype in a case population is compared to frequency of the haplotype in the control population.

6. Method according to any one of the preceding claims wherein the haplotypes are deduced over a scan window of at least 10 kb.

7. Method according to any one of the preceding claims wherein less than 20% of the individuals in the sample population are a first degree relative of another individual in the sample population.

8. Method according to any one of the preceding claims wherein the phenotype is a disease or response to medication.

Fig 1 SNPs of the *Nat2 Gene*

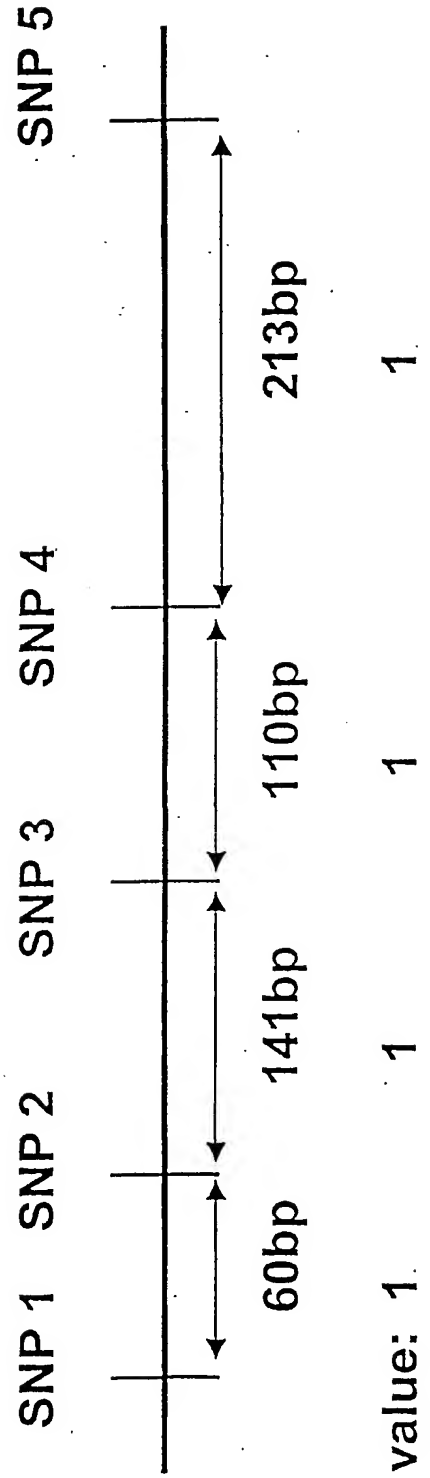

Fig 2 SNPs over the Chr X Locus